Support Verb Constructions across the Ocean Sea

Jorge Baptista\textsuperscript{1,2}, Nuno Mamede\textsuperscript{2,3} & Sónia Reis\textsuperscript{1,2}

\textsuperscript{1}University of Algarve, Campus de Gambelas, 8005-139 Faro, Portugal
\textsuperscript{2}INESC-ID Lisboa - Human Language Technology Lab, Alves Redol 9, 1000-029 Lisboa
\textsuperscript{3}Instituto Superior Técnico, University of Lisbon, Av. Rovisco Pais 1, 1049-001 Lisboa

jbaptis@ualg.pt, nuno.mamede@tecnico.ulisboa.pt, smreis@ualg.pt
abstract

This paper analyses the support (or light) verb constructions (SVC) in a publicly available, manually annotated corpus of multiword expressions (MWE) in Brazilian Portuguese. The paper highlights several issues in the linguistic definitions therein adopted for these types of MWE, and reports the results from applying STRING, a rule-based parsing system, originally developed for European Portuguese, to this corpus from Brazilian Portuguese.

The goal is two-fold: to improve the linguistic definition of SVC in the annotation task, as well as to gauge the major difficulties found when transposing linguistic resources between these two varieties of the same language.

Keywords: support-verb constructions, light verbs, predicate noun, Portuguese, Brazilian/European Portuguese
summary

**introduction:** concept of Support Verb Constructions (SVC), definitory properties, goals

**related work:**

**SVC in PARSEME:** LVC.full/cause, corpus compilation/annotation, lexical variety; linking operator-verb (Vopl), causative operator-verb (Vopc), standard/converse SVC

**processing SVC** in PARSEME corpus through STRING

**results and discussion:** FN: on-going/missing, missing Vsup, parsing issues;
TN: PT-BR exclusive SVC; FP: parsing issues

**conclusion and future work**
introduction

Support verb constructions (SVC)

- fundamental component of many languages (universal?)
- large variety of semantic predicates (like predicative verbs and adjectives)
- support verb + predicate noun + argument(s) (M. Gross 1981)
- Vsup serves basically to “conjugate” the predicate noun (M. Gross 1989, p.38)
- Vsup conveys grammatical (tense, aspect) and stylistic values (M. Gross 1998)

(terminology note: support verb = light verb; Fotopoulou et al. 2021)
introduction (cont.)

(1) O **Pedro** *tem* fome
   lit: ‘Pedro has hunger’ ‘Pedro is hungry’ (Santos 2015)

(2) O **Pedro** *deu* um abraço ao **João**
   ‘Pedro gave a hug to João’ (Baptista 1997b; Calcia 2022)

(3) O **Pedro** *fez/está_em* greve
   ‘Pedro is on strike’ (Chacoto 2005; Dias de Barros 2014; Ranchhod 1990)

(4) É de _o interesse do **Pedro** que o **João** faça isso
   ‘It is in Pedro’s interest that João do this’ (Baptista 2005).
Support verb constructions (SVC)

- SVC properties defined by Vsup+Npred combination
- Regularities can be found but can not be generalized over the lexicon of neither Vsup nor Npred
- Definitory properties, distinguishing SVC from ordinary verb constructions e.g. argument selection:

(5) *O Pedro_i deu um abraço_i (*meu/*do Rui_k) ao João
    ‘Pedro gave a/my/Rui’s hug to João’
This paper’s goals:

- assess (on-going) SVC lexicon-grammar description of (European) Portuguese (PT-PT) (see references)
- assess its implementation in STRING, a hybrid, rule-based and statistical natural language processing pipeline for PT (Mamede et al. 2012, Baptista & Mamede 2020)
- use a publicly available, independently annotated, corpus (PARSEME) built for Brazilian Portuguese (PT-BR), with SVC manually annotated
- discuss several issues in the linguistic definition of SVC adopted in this corpus
- gauge the major difficulties when applying a lexicon-grammar built for PT-PT to a corpus of PT-BR, highlighting the similarity of the two major varieties of Portuguese
related work

○ quite old, the idea of nouns as predicates (Port-Royal, Arnaud & Lancelot 1660)
○ the terms support verb and predicative noun coined by M. Gross (1981)
○ concepts later revised and extended (M. Gross 1998)
○ extensive/systematic descriptions of SVC within the Lexicon-Grammar framework for major European and other non-IE languages (M. Gross 1996)
○ European Portuguese (early 80’s and 90’s); Brazilian Portuguese (2013-2022)
○ recent and brief, non-exhaustive overview on the topic (Fotopoulou et al. 2021)
related work (cont.)

- **CVS in Portuguese** (overview)
  - **European Portuguese**
    - specific constructions and/or transformations: conversion (Baptista 1997a; Baptista 1997b, Baptista 2005a); fusion and instrument nouns (Baptista 2004); symmetry (Baptista 2005b), communication predicates (Reis et al. 2021), non-agentive SVC (Baptista et al. 2021)
  - **Brazilian Portuguese**
    - from **2010s** (Dias de Barros 2014: *fazer*; Rassi 2015: *dar* ‘give'; Santos 2015: *ter* ‘have’)
    - specific constructions and/or transformations: conversion (Calcia 2016; Calcia and Vale 2019; Calcia 2022), aspectual variants of vsup (Picoli et al. 2021), non-agentive SVC (Dias de Barros et al. 2013)
related work (cont.)

- Few works systematic comparison of the lexicon and grammar of the PT-PT and PT-BR variants (Rassi et al. 2016: converse SVC)
- annotated corpus of SVC with support-verb *dar* 'give' (Rassi et al. 2015b)
- processing SVC in Portuguese (Baptista et al. 2015; Rassi et al. 2014; Rassi et al. 2015a)
- recent developments in (Mota et al. 2018; Baptista & Mamede 2020; Barreiro et al. 2022)
- extensive literature on SVC across multiple languages, on their place within the description of multiword expressions (Sag et al. 2002); relation of SVC with fixed, verbal idioms; challenges SVC pose to Natural Language Processing (NLP) (Constant et al. 2017, for an overview)
- significant boost multilingual community COST project PARSEME (Savary et al. 2015)
related work (cont.)

PARSEME project

- goal “characterizing MWEs in lexicons, grammars and corpora and enabling systems to process them” (Ramisch et al. 2020, p.107)
- shared task on automatic identification of MWE (Ramisch et al. 2020)
- (Brazilian) Portuguese corpus, manually annotated for verbal MWE
- “unified guidelines for all the participating languages, in order to avoid heterogeneous, hence incomparable, datasets”: decision-trees
- main categories: SVC (called LVC=light verb) and verbal idioms
SVC in PARSEME

Some data (from documentation)*:

- 80% training + 10% development + 10% testing partitions
- total: 27,9 K sentences, 638 K tokens
- 3,145 LVC.full (= SVC) and 94 LVC.cause (= causative constructions)
- testing partition: 2,770 sentences, +62.6 K tokens, 337 LVC.full, 7 LVC.cause
- “sentences from the informal Brazilian newspaper Diário Gaúcho and from the training set of the [Universal Dependencies] treebank” (UD_Portuguese-GSD v2.1); no further information found on the sampling/selection method for the sentences included in the corpus
- 1,000 sentences doubly-annotated; agreement reported seems to indicate adequate guidelines, yielding reasonable consistency, given the complexity of the task

* Ramisch et al. (2020); https://gitlab.com/parseme/parseme_corpus_pt
Lexical variety
(for the purpose of this paper)

football (soccer) domain vocabulary:

- 134 instances of *gol* 'goal' (*fazer* 'do' and *marcar* 'score', *dar* 'give');
- 30 *falta* (*fazer*, *marcar*, *sofrer* 'suffer', *cometer* 'comit');
- 13 *passe* (*fazer* 'make, do', *receber* 'receive');
- 11 *pênalti* 'penalty' (*marcar*, *sofrer*); etc.

6.8% SVC instances concern vocabulary of this domain
Predicate nouns designating **measurable quantities**:

8 área ‘area’, 2 população ‘population’; some taken from the Astronomy domain: 2 ascenção reta ‘right ascension’, 2 declinação ‘declination’, 6 excentricidade ‘eccentricity’, inclinação ‘inclination’:

*Possui uma excentricidade de 0.03574140 e uma inclinação de 11.03095°.*

‘It has an eccentricity of 0.03574140 and a tilt of 11.03095°.’
SVC in PARSEME (cont.)

- Predicate nouns designating **measurable quantities**: 
  
  *Possui uma excentricidade* de 0.03574140 e *uma inclinação* de 11.03095º. 
  ‘It has an eccentricity of 0.03574140 and a tilt of 11.03095º.’

- Apparently, no occurrence of other SVC of some of these nouns (Baptista 2005): 
  
  *O Pedro é de uma certa excentricidade* 
  lit: ‘Pedro is of a certain eccentricity’ (‘Pedro is eccentric’) 
  *O Pedro tem uma certa inclinação para a música* 
  lit: ‘Pedro has a certain inclination to music (‘Pedro ’)

- Nor any case of **Restructuring** (Baptista & Ranchhod, 1998): 
  
  *O mundo tem 510 bilhões de km² de área total* (www.ufjf.br) 
  ‘The world has 510 billion km² of total area’
SVC in PARSEME (cont.)

- Linking operator verbs (M. Gross 1981)

*A Cátedra Milton Santos* tem como objetivo a difusão de informações (id=pt br-ud-train-s7942)

‘The Milton Santos Chair has as its aim/purpose to disseminate information’

*o especial da TV Globo* terá como tema a vida de Dolores Duran (id=diario gaucho 16311)

‘the TV Globo special will have as theme the life of Dolores Duran’
SVC in PARSEME (cont.)

- there is an underlying SVC, with standard syntactic configuration, and the predicative natures of *como* hints at the sentence with copula verb:

A (constituição da) Cátedra Milton Santos *tem um certo objetivo* 
# esse objetivo é a difusão de informações
‘The (constitution of the) Milton Santos Chair has as an aim/purpose 
# This aim/purpose is to disseminate information’

o especial da TV Globo *terá um certo tema* # o tema é a vida de Dolores Duran
‘the TV Globo special will have a theme # This theme is the life of Dolores Duran’
SVC in PARSEME (cont.)

- Linking operator verbs (continued), on adjectival/participial constructions:

  Já Federer . . . teve uma campanha mais perturbada ...
  ‘Federer [=person] ... had a more troubled/disrupted campaign’

  cp. a campanha (de Federer) foi perturbada  ‘the campaign was troubled/disrupted’
SVC in PARSEME (cont.)

- Linking operator verbs (continued), on prepositional constructions:

  \[ \text{Pelo segundo ano consecutivo, o Cruzeiro teve uma campanha abaixo de\_as expectativas.} \]
  
  (lit.: had a campaign below expectations’)
  
  ‘For the second year in a row, Cruzeiro [football club] has fallen short of expectations’
  
  cp. \[ \text{a campanha (do Cruzeiro) esteve abaixo de\_as expectativas} \]
  
  ‘the campaign was below expectations’
SVC in PARSEME (cont.)

- A clear SVC construction of the noun **campanha** ‘campaign’ is patent in the corpus:

  Os jovens...**estão fazendo a campanha** com a cara e a linguagem deles
  ‘Young people are making the campaign with their face and their language’

- some cases where the notion of support seems too much stretched:

  . . . o partido foi vítima de uma intensa **campanha promovida pela oposição de direita e seus aliados** . . . ‘the party was the victim of an intense campaign promoted by the right-wing opposition and its allies’

  cp. **O Pedro promoveu a campanha** (**do João** = **que o João fazia**)
  ‘Pedro promoted the campaign (of João/João’s = that João was doing)’

- **ser vítima de Npred** and SVC with copula (see below)
Causative operator verbs (Vopc; M. Gross 1981), construct integrated in PARSEME Guidelines (LVC.cause)

*A ausência do sexo também traz uma forte angústia* (OK: marked as LVC.cause)
‘The absence of sex also brings a strong anguish’

*Nós ... estamos ansiosos para montar um time competitivo, que seja divertido e traga orgulho para os fãs* (Vopc was missed)
‘We are looking forward to building a competitive team that is fun and brings pride to the fans’

*Caro V., a tua postura é sempre admirável, o que faz com que tua lealdade a esta coluna só me dé orgulho* (one Vopc missed and another marked as LVC.full)
‘Dear V., your posture is always admirable, which makes me proud of your loyalty to this column’
SVC in PARSEME (cont.)

operator-verbs (conclusion)

- distinction between \textit{Vsup} and \textit{Vopc} construction is not consensual
  - (e.g. PT-PT: de Athayde 2000 vs. Chacoto 2005)
  - this paper concurs with PARSEME classification criteria:
    semantic neutral value of \textit{Vsup} vs. semantic added value of \textbf{cause} of \textit{Vopc}

- theoretical construct of linking operator-verb
  - also proposed by M. Gross (1981), but largely ignored in the literature
  - treated as a ordinary \textit{Vsup} in PARSEME (or just ignored)
  - no new meaning element is introduced, came predicate noun (and partially its structure)
  - proposed as a \textit{saliency-inducing device} (Ranchhod 1990, 183 ff.)
  - a new notation is required, distinct from \textit{Vsup} and \textit{Vopc}
SVC in PARSEME (cont.)

- SVC with *ser* and *estar Prep* (‘be’ and ‘be Prep’)
  - ignored (or/and eventually excluded from SVC) as per PARSEME Guidelines
  - (very few) of this SVC expressions can be found in the corpus

<Title of a song> é sucesso!
‘<title of song> is [a] hit (lit.: success)’

. . . um comercial da Tim *era sucesso na tevê*  ‘a Tim’s commercial was a hit on TV’

Você *está com saudade* do frio?
‘Do you miss the cold?’

*Ela está com febre, dores no corpo e coriza* (id=diario gaucho 16030)
‘She has (lit.: is with) a fever, body aches and a runny nose’;

*Quando não* *estão de licença* (id=pt br-ud-train-s7822)
‘When they are not on leave of absence’
SVC in PARSEME (cont.)

- **Standard/Converse SVC; [Conversion]** (M. Gross 1981; G. Gross 1989)

(ST)  
O Pedro *deu* um abraço ao João  

= (CV)  
O João *recebeu/levou* um abraço do Pedro  

- ‘Peter gave John a hug’
- ‘John got a hug from Peter’

- very productive transformation, occur in several (many) languages
- extensive descriptions in PT-PT and PT-BR (Baptista 1997a, 1997b, 2005b; Rassi *et al.* 2016; Calcia 2016; Calcia & Vale 2019; Calcia 2022).
- occurred often in the corpus (34 times)
- may have an impact in the semantic role labelling
- since PARSEME ignored this distinction, this Standard/Converse opposition was also ignored
processing SVC in PARSEME corpus with

- STRING pipeline (Mamede et al. 2012; Baptista & Mamede 2020)
  - basic text processing tasks, including text segmentation, tokenization, dictionary-based part-of-speech (PoS) tagging, rule-based and statistical PoS disambiguation, and parsing.
  - XIP Parser (Aıt-Mokhtar et al. 2002) with the STRING Portuguese lexicons and grammar
  - Dependency relations: SUBJ[ect], CDIR (direct object), etc., between chunks heads
  - SVCs are formally identical to ordinary verbal constructions > STRING overall strategy
    - firstly, capturing (general) syntactic dependencies between predicate noun and the support verb; and
    - then extracting a specific dependency SUPPORT linking the predicate noun and the support verb.
  - The system can be configured to output only the desired dependencies.
processing SVC in PARSEME corpus with STRING (cont.)

A Ana *marcou* dois *gols*  ‘Ana has scored (lit.: marked) two goals’

>TOP{NP{A Ana} VF{marcou} NP{dois gols}}

SUBJ_PRE(marcou,Ana)

CDIR_POST(marcou,gols)

SUPPORT_VSUP-STANDARD(gols,marcou)
processing SVC in PARSEME corpus with STRING (cont.)

if ( ( VDOMAIN(#1,#2[lemma:"fazer"])) || VDOMAIN(#1,#2[lemma:"marcar"])) 
   & (MOD[post,relat](#3[lemma:"golo"],#2)) ||
      CDIR(#2[transf-passiva:~],#3[lemma:"golo"])) ||
      SUBJ(#2[transf-passiva],#3[lemma:"golo"])) ||
      (ANTECEDENT[relat](#3[lemma:"golo"],#4[pronrel])
         & SUBJ(#2[transf-passiva],#4))
   & ~ SUPPORT[vsup-standard](#3,#2) )
SUPPORT[vsup-standard=+](#3,#2)

- **15,615 rules** automatically generated from the database
  (lexicon-grammar; Baptista & Mamede 2020, *updated*)
- **5,800 SVCs** (+3,320 predicate nouns, *ongoing*);
- gauge lexical coverage of the resource and the system’s performance on a previously unseen corpus, from a different variety
processing SVC in PARSEME corpus with STRING (cont.)

- (in-house script) retrieve {npred, vsup} pairs linked by LVC.full dependency from corpus
- corpus processed by STRING, (only) with SUPPORT dependency output: (2)
- convert LVC.full into SUPPORT and merge with processed corpus output (1)
- semi-automatic comparison
processing SVC in PARSEME corpus with STRING (cont.)

1>TOP{PP{PP{Em 2} PP{de outubro} PP{de 2009}} , PP{ em o 1º aniversário} PP{de a SmackDown} , NP{The Rock} VF{fez} NP{uma aparição} AP{especial} PP{em um vídeo} AP{pré-gravado} .

   PARSEME: SUPPORT_VSUP-STANDARD(aparição,fez)
   STRING: SUPPORT_VSUP-STANDARD(aparição,fez) (true-positive)

2>TOP{NP{Se} NP{nenhuma providência} VCOP{for} VCPART {tomada} , NP{a população} VTEMP{vai} VASP{voltar a} VINF{usar} NP{lamparinas} ADVP{ADV{a a noite}} e NP{geladeira} PP{a querosene} .

   PARSEME: SUPPORT_VSUP-STANDARD(providência,tomada)
   STRING: -- (false-negative)
results and discussion

- 2,770 sentences (as segmented by STRING)
- 311 instances of LVC.full in PARSEME
- high precision / low recall

Error analysis

- false-negatives:
  - ongoing predicate nouns’ description
  - missing nouns
  - missing support verb in the lexicon-grammar
  - pos-tagging and parsing issues

- true-negatives
  - predicate nouns exclusive of PT-BR
  - lexical variants exclusive of PT-BR
  - spelling errors/variants

- false-positives
  - pos-tagging and parsing issues

<table>
<thead>
<tr>
<th>TP</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>197</td>
<td>20</td>
<td>270</td>
</tr>
</tbody>
</table>

\[ P = 0.91 \]
\[ R = 0.42 \]
\[ F = 0.58 \]
results and discussion (cont.)

- ongoing predicate nouns’ description

- missing nouns
missing support verb in the predicate noun entry of the lexicon-grammar


converse support-verbs

results and discussion (cont.)

- parsing issues
  - error in PoS-disambiguation: ajuste (noun/verb)

  A prática de fazer ajuste no superávit com os dividendos tem sido comum nos últimos anos
  ‘The practice of making adjustment the surplus with dividends has been common in recent years’

- particular syntactic constructions not captured by the grammar

  Três integrantes de um bando que fez um dos maiores ataques a banco dos últimos anos
  ‘Three members of a gang that carried out one of the biggest bank attacks in recent years’

  complex partitive determiner construction (M. Gross 1977):
  fez um [ataque (que foi)] dos maiores ataques
  only partially parsed (no CDIR): QUANTD_PRE(ataques,um)
false-positives

*Seria uma boa surpresa e uma prova de que amor não tem hora nem dia marcados.*

‘It would be a nice surprise and proof that love has no set time or day.’

**STRING: que** (PoS-tagging error: *Pron-rel vs. Conj-sub)

INTROD_RELAT(que,tem), ANTECEDENT_RELAT(prova,que), MOD_POST_RELAT(prova,tem)

→ SUPPORT_VSUP-STANDARD(prova,tem)

SVC and the last stages of parsing, accumulation errors (PoS-tagging, parsing)
This paper:
- analyzed the support-verb constructions (SVC) manually and independently annotated in a publicly available corpus of Brazilian Portuguese (PT-BR), originally built within the scope of the project PARSEME
- corpus was parsed with STRING NLP pipeline for Portuguese
- using PT-PT SVC lexicon-grammar (still on construction: 2/3)
- described SVC found in corpus
- compared SVC dependencies annotated in corpus with STRING system output
- high precision (.91) but low recall (.42), still much room for improvement
- error analysis: on-going (⅔), PT-BR nouns, PT-BR-specific or missing support-verb+noun combinations (e.g. possuir ‘possess’, passar ‘pass’), PoS-tagging and parsing errors; still rule improvement
- (partition of) corpus with reference output available at:
Thank you!
references


references


